

# Measuring Teaching Matters

## What Different Ways of Looking at Student Results Tell Us About Teacher Effectiveness

July 2013

States are redesigning their educator effectiveness systems to provide more information and more support to improve teaching. In the process, they increasingly look beyond the most basic and historically most common view of measuring student performance: how many students “passed” the State test in a given year. A number of States now require an objective measure of student growth to be part of teacher evaluations.<sup>1</sup>

States are using student growth measures to understand teacher effectiveness for good reasons. First, student learning is the most important expectation we set for schools, and nothing in a school impacts student learning more than effective teaching.<sup>2</sup>

Second, new data systems permit far better links between student outcomes (tests, graduation, postsecondary experiences) and specific schools and teachers. This facilitates assessment of and systemic learning about changes to policy and practice that might lead to improvements in the quality of teaching and public schools.

Finally, traditional methods of evaluating teachers that typically do not include objective measures of teacher performance have in most state education agencies (SEAs) and local educational agencies (LEAs) provided inadequate information about teacher effectiveness. In particular, these methods tend to yield high ratings for almost all teachers, and consequently these ratings have little value in predicting either future teacher effectiveness or student achievement.<sup>3</sup> As a result, they have yielded little information that can help teachers become more effective practitioners.

This brief describes various approaches to measuring student growth and what research says about the extent to which student growth may be used as a measure of teacher performance.

### Approaches to Measuring Teacher Performance

There are a number of different methods SEAs and LEAs can use to translate student test achievement into a measure of teacher performance. All of these involve predicting how well students in a teacher’s classroom will perform on tests and contrasting this to their actual performance. The main difference

between the statistical methods described below is the way in which researchers, SEAs and LEAs form expectations or predictions about student achievement. Once they create predictions about how students will do, they attribute the aggregated difference (across students in a classroom) between these predictions and actual student achievement, at least to some degree, to teachers. This forms the basis of a teacher performance measure.

An increasing amount of literature explores the extent to which different methods of predicting student achievement and then aggregating these across students in a classroom results in accurate measures of teacher performance.<sup>4</sup> This brief touches on this issue, but its primary purpose is to describe different models, not assess their validity.

In general, models that take into account where students start before some educational intervention or receiving instruction from a particular teacher add an important dimension to understanding learning and the contributions that schools and teachers make to learning.<sup>5</sup> Most of this brief offers an overview of some of the more common of these models. It concludes with a discussion of the importance of using multiple measures to identify teacher performance.<sup>6</sup>

## Value-Added Models

Value-added models (VAMs) are a class of models that measure student test achievement against some prediction of how students are expected to do given their earlier achievement level and, depending on the specific model, other factors thought to both influence student learning and reside outside the control of teachers and schools.<sup>7</sup> Educational researchers have long used the value-added framework to address questions about the efficacy of different interventions and the effects of different levels of school resources, such as class size.<sup>8</sup>

SEAs and LEAs can use VAMs to help answer questions about actual performance in light of expected results: Is the learning this student demonstrated on this year's test greater or less than would be expected, in light of the performance of other students with similar prior achievement and similar backgrounds?

As part of teacher evaluation systems, VAMs aim to predict what student growth can be expected from an average or typical teacher, and then compare actual student achievement with that prediction. A teacher's value-added score is intended to convey how much individual teachers contribute to student learning in a particular subject in a particular year. Teachers who

produce more than this typical teacher are thought to have added value. Teachers whose effects on students result in less growth than the typical teacher is expected to yield are considered less effective.

VAM measures of teacher performance differ according to the particular VAM used because models differ in terms of how they adjust for student and out-of-school factors that influence achievement and the way in which they compare teachers. Some models, for instance, predict only student achievement based on prior test scores, while others include controls for factors such as a student's race and ethnicity, eligibility for free or reduced-price lunch, and so on. And teacher performance, for example, may be judged relative to other teachers in the same school or relative to a larger set of teachers, such as those in a LEA or a whole state. The differences between models are sometimes small, but can also have meaningful impacts on estimates of teacher performance, particularly for teachers who are serving students with backgrounds that differ from those in an average classroom.<sup>9</sup>

## Gain Score Model

Many educators and policymakers are familiar with one VAM, the Gain Score model. This model measures changes in individual student achievement between two or more points in time, for example from the beginning to end of a school year or from one administration of an annual test to the next, but does not include any statistical adjustments for the type of students served or the resources that schools or teachers have on hand.

A virtue of gain scores is that they are easy to calculate in that they simply entail taking achievement for an individual student in a particular grade and subject and subtracting, for instance, the score from achievement in the same subject in the prior grade. Measuring student growth in this way is not new. Teachers and researchers have long used pre- and post-test designs to understand learning over time and gauge how much new knowledge students gain as a consequence of, for instance, an intervention or classroom lesson. This model helps answer questions about the amount

of learning that has taken place: How much did this student learn last year? How much, on average, did this teacher's students' performance change over the course of this school year? How much did this group of eighth-graders learn compared with all eighth-graders in the State?

To explore the progress made by students in a particular classroom assigned to a specific teacher, a growth/gain model typically averages the gains across the class and compares that average with those of other teachers. Were this average gain used as a measure of teacher performance, the implicit prediction assumption is that students would all have equal achievement gains in the absence of differences among teachers: Differences in average gains represent what is attributed to teachers.

### Student-Growth Percentile Model

Users of the Student Growth Percentile (SGP) model want it to facilitate a comparison of learning from one grade to the next or one test to the next when the desire is to assess how student performance compared with other students with a similar prior achievement level. The SGP model uses a statistical procedure called "quantile regression" that calculates where in the achievement distribution a student falls relative to other students with a similar prior test score history.<sup>10</sup> Thus, for example, a student who has a growth percentile of 75 in the 4th grade had test achievement growth from the 3rd to the 4th grade that equaled or exceeded 75 percent of *students who started with a similar prior achievement level* in the 3rd grade.

The SGP model can answer certain questions, such as: Did this student learn as much this year as she learned last year? Did this student learn as much in math as he learned in reading? Did these students learn as much as their peers in our LEA? Which program or instructional approach resulted in the most student learning?

When examining results aggregated across students, it is typical to calculate an average or median of all the student growth percentiles at the level of interest—for instance, a classroom, school or LEA. Using the SGP

model, States also can set a target for adequate growth, for example, the growth needed to reach or maintain proficiency.

### Value Added: What the Research Says

Since Tennessee pioneered use of a VAM in its State accountability system in the early 1990s,<sup>11</sup> student growth models and VAMs have become more refined and sophisticated. The use of these models to gauge learning and the impact of specific teachers continues to develop and evolve. A growing body of research shows that this kind of analysis contributes valuable information to understanding teaching and learning.

### Information about student achievement and teacher contributions to student results

VAMs do not predict or attribute student achievement with pinpoint accuracy because not everything about student achievement can be predicted easily and explained with the data that is typically at hand. Still, research has found that value-added measures predict future student achievement better than other factors commonly used for important personnel decisions about teachers.<sup>12</sup> Used with other measures, VAMs can increase understanding about teacher practice and its connection to student learning.<sup>13</sup>

Research has found that these models hold much promise in using student test results to make reasonably reliable inferences about teacher effectiveness.<sup>14</sup> Moreover, the results from VAMs match closely principals' evaluations of the most effective and least effective teachers.<sup>15</sup>

### A role for multiple measures of teacher effectiveness

Although research shows that growth models and VAMs shed important and unique light on teacher effectiveness, studies also suggest that SEAs and LEAs should be cautious about basing high-stakes decisions solely on these models. Researchers and practitioners still disagree about whether and how student test-based measures of teacher performance should be

used. In addition, test-based measures alone do not provide teachers with timely feedback on their practices.

SEAs and LEAs should consider a number of limitations to VAMs. The statistical techniques may not account fully for the fact that students are generally not randomly assigned to teachers. Without documenting and modeling the particular ways that schools distribute teachers in classrooms, some experts argue it is difficult to fully account for how those choices affect the effects attributed to each teacher's skills, efforts and success.<sup>16</sup>

The realities of testing—students guess both correctly and incorrectly, students come to school with the flu on testing day—could also fog the picture of effectiveness painted by the results and lead to the inaccurate classification of teachers. Variation in the value teachers are found to add from year to year also has raised questions about the models. This leaves some wondering if it is plausible for teacher effectiveness to vary so greatly. The Brown Center on Education Policy at Brookings has noted, however, that these year-to-year variations are consistent with annual job performance measures in other fields. They also resemble the variations between SAT scores and first-year college grade-point averages—and most feel comfortable about using SAT scores to make high-stakes decisions about students' college readiness.<sup>17</sup> And, thus, while it is vital to understand limitations and to use results carefully, growth models and VAMs improve on teacher effectiveness systems that already are notoriously inaccurate: SEAs and LEAs have routinely deemed all but a relative handful of teachers to be effective.<sup>18</sup>

LEAs and SEAs are overcoming these challenges as VAMs grow more common and more sophisticated. Using several years of data may make the estimates of teacher effectiveness more stable; the models' reliability increases with additional years of data (particularly up to three years of results).<sup>19</sup> The data systems that LEAs and States rely on are also getting better at identifying which students are assigned to which teachers for different types of instruction.

Still, given some of the limitations of student test-based measures of teacher effectiveness, there is an emerging consensus that SEAs and LEAs would be well-served to use multiple measures to arrive at a summative assessment of teacher performance. Even proponents who contend that "value-added is superior to other existing methods of classifying teachers,"<sup>20</sup> suggest that teacher evaluation should have many facets.

The Measures of Effective Teaching (MET) project funded by the Bill and Melinda Gates Foundation recommends that classroom observations, student achievement gains and student survey feedback be used *together* as a set of multiple measures to evaluate teachers. MET researchers hold that the combination of these multiple measures increases the ability to predict future student achievement, improves reliability, and provides richer diagnostic feedback that teachers can use to improve. The MET project has found that teachers who demonstrated greater effectiveness in classroom observations had higher student achievement gains than other teachers. However, classroom observations alone did not predict student achievement as reliably as observations, combined with student feedback and achievement gains.

## Conclusion

As SEAs and LEAs introduce more comprehensive educator effectiveness systems that include measures of growth in learning and value added, they face many technical issues. Beyond the challenge of measuring growth in non-tested grades and subjects, States must address the quality of the assessments they use, including whether they can link test results vertically, from grade to grade. States must contend with issues of awareness and understanding as well. The more complex questions they seek to answer, the more complex and less transparent the model for arriving at the answers become.

Although many laypeople can relatively easily calculate gain-scores, it is more complex for teachers or the public to replicate the results from value-added or SGP models given the statistical expertise required to do so, the size of the data sets they marshal and

concerns about student privacy when exploring how academic peers perform. States are well served to go to great lengths to explain their models and to make them as transparent as possible.

Shifting to growth models and VAMs is a significant step SEAs and LEAs have taken to improve their data systems in recent years. SEAs' work to build high-quality data, longitudinal capacity and the ability to match students and teachers to each other and to student results all make the use of the models more feasible.<sup>21</sup>

VAMs and growth models are new and rely on data systems that SEAs and LEAs are still building, but preliminary research shows that "combining new approaches to measuring effective teaching—while not perfect—significantly outperforms traditional measures."<sup>22</sup> Used as part of a body of evidence collected to measure student growth and teacher effectiveness, VAMs and growth models hold promise for helping policymakers and practitioners collect and analyze sophisticated data on teaching and learning that can guide professional development, the distribution of human resources and decisions about career milestones within the nation's schools.

Yet SEAs and LEAs can use value-added approaches only when they consistently use standardized tests across several grades and subjects. These conditions exist in most States testing third-grade through eighth-grade mathematics and reading/English Language Arts, and high school mathematics and English language arts. But in most LEAs, testing programs provide data to measure the effectiveness of less than half of the teachers. Measuring learning in non-tested grades and subjects presents challenges to using growth models and VAMs. The Reform Support Network has begun to address these challenges through a seminar on measuring student growth in non-tested grades and subjects, a student learning objective work group of Race to the Top states and other publications, including a guide on student learning objectives.

## Endnotes

<sup>1</sup> See National Council on Teacher Quality, *State Teacher Policy Yearbook 2011*, [http://www.nctq.org/stpy11/reports/stpy11\\_national\\_report.pdf](http://www.nctq.org/stpy11/reports/stpy11_national_report.pdf).

<sup>2</sup> Erik A. Hanushek, "The Trade Off Between Child Quantity and Quality," *Journal of Political Economy* 100(1) (1992): 84-117; Erik A. Hanushek and Steven G. Rivkin, "Generalizations About Using Value-Added Measures of Teacher Quality," *American Economic Review* 100(2) (2010): 267-71; Steven G. Rivkin, Erik A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica*, 73(2) (2005): 417-458; Jonah Rockoff, "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review* 94(2) (2004): 247-252; William L. Sanders and Sandra P. Horn, "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research," *Journal of Personnel Evaluation in Education* 12(3) (1998): 247-256; Robert Gordon, Thomas J. Kane, and Douglas O. Staiger, "Identifying Effective Teachers Using Performance on the Job" (Hamilton Project Discussion Paper, The Brookings Institution, 2006).

<sup>3</sup> See Daniel Weisberg, Susan Sexton, Jennifer Mulhern, and David Keeling, "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness," (New York: The New Teacher Project, 2009), <http://widgeteffect.org>.

<sup>4</sup> See, for instance, Raj Chetty, John N. Friedman, and Jonah Rockoff, "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood" (National Bureau of Economic Research working paper no. 17699, 2011); Daniel F. McCaffrey, J.R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton, Laura S, "Evaluating Value-Added Models for Teacher Accountability" (Santa Monica, California: RAND Corporation, 2003); Thomas J. Kane, Daniel F. McCaffrey, T. Miller, and Douglas O. Staiger, "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignments" (Seattle, Washington: Measures of Effective Teaching Project, Bill & Melinda Gates Foundation, 2013); and Jesse Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125(1) (2010): 175-214.

<sup>5</sup> In particular, such models contrast with what is commonly referred to as the *achievement or status model*, which is a snapshot showing how much a student knows compared with a standard or target. The model measures student achievement at a point in time without consideration of students' achievement when they entered a school or classroom or any other factors. It can answer the most direct questions about where an individual or group of students (for example, in a classroom or school) stands academically relative to others. However, it does not measure how individual students progress over time, and thus the amount a student has learned when enrolled in a particular school or classroom. Given this, the achievement model is considered to be a very poor measure of teacher performance.

- <sup>6</sup> Readers wanting a more in-depth explanation of various statistical models might see Dan Goldhaber, Brian Gabele, and Joe Walch, "Does the Model Matter? Exploring the Relationship Between Different Achievement-Based Teacher Assessments" (CEDR Working Paper 2012-6, University of Washington, Seattle, 2012); Douglas Harris, "Value-Added Measures in Education: What Every Educator Needs to Know" (Cambridge, Massachusetts: Harvard Education Press, 2011); and Daniel F. McCaffrey, J.R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton, "Evaluating Value-Added Models for Teacher Accountability" (Santa Monica, California: RAND Corporation, 2003)
- <sup>7</sup> These factors include a student's poverty level or the spending level at a school.
- <sup>8</sup> See Eric Hanushek, "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, 14(3) (1979): 351-388.
- <sup>9</sup> See, for instance, a discussion about this in Goldhaber et al. "Does the Model Matter?" (CEDR Working Paper 2012-6, University of Washington, Seattle, 2012), at <https://appam.confex.com/appam/2012/webprogram/Paper2264.html>.
- <sup>10</sup> Student growth percentiles conceptually resemble the height and weight percentiles commonly used by pediatricians to chart growth and inform parents as to children's physical development. The percentiles show how students compare with other students in the distribution of performance.
- <sup>11</sup> Sanders and Horn, "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database," *Journal of Personnel Evaluation in Education* 8(1) (1994): 299-311. For more information on TVAAS, see [http://www.tn.gov/education/assessment/doc/TVAAS\\_Fact\\_Sheet.pdf](http://www.tn.gov/education/assessment/doc/TVAAS_Fact_Sheet.pdf).
- <sup>12</sup> Dan Goldhaber and Michael Hansen, "Using Performance on the Job to Inform Teacher Tenure Decisions," *American Economic Review* 100(2) (2010):250-255 at <http://www.urban.org/publications/1001385.html>.
- <sup>13</sup> For a discussion on this see, for instance, "Feedback for Better Teaching: Nine Principles for Using Measures of Effective Teaching." (Seattle, Wash.: The Measures of Teaching Project, Bill and Melinda Gates Foundation, January 2013) at [http://www.metproject.org/downloads/MET\\_Feedback%20for%20Better%20Teaching\\_Principles%20Paper.pdf](http://www.metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles%20Paper.pdf).
- <sup>14</sup> Glazerman et al "Evaluating Teachers: The Important Role of Value-Added." Washington, D.C.: Brown Center on Education Policy at the Brookings Institution, 2010.
- <sup>15</sup> Brian Jacob and Lars Lefgren, "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics* 26(1) (2008): 101-136 at <https://economics.byu.edu/Documents/Lars%20Lefgren/papers/principals.pdf>;" Douglas N. Harris and Tim R. Sass, "What Makes for a Good Teacher and Who Can Tell?" (CALDER Working Paper # 30, 2009), at <http://www.urban.org/publications/1001431.html>.
- <sup>16</sup> See Henry Braun, "Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models" (Princeton, New Jersey: Educational Testing Service. Economic Policy Institute, 2010); "Problems with the Use of Student Test Scores to Evaluate Teachers" (EPI Briefing Paper); and, also see Jesse Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125(1) (2010): 175-214.
- <sup>17</sup> Steven Glazerman, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, and Grover Whitehurst, "Evaluating Teachers: The Important Role of Value-Added" (Brown Center on Education Policy, Brookings Institution, 2010), <http://www.brookings.edu/research/reports/2010/11/17-evaluating-teachers>.
- <sup>18</sup> See Weisberg et al., "The Widget Effect" (The New Teacher Project, 2009) at <http://widgeteffect.org/downloads/TheWidgetEffect.pdf>.
- <sup>19</sup> Dan Goldhaber and Michael Hansen, "Is it Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance," *Economica* (forthcoming).
- <sup>20</sup> See Glazerman et al., "Evaluating Teachers" (Brookings Institution).
- <sup>21</sup> See "Data Quality Campaign" for information on State data system capacity, <http://www.dataqualitycampaign.org>
- <sup>22</sup> Thomas Kaine and Douglas Staiger, "Gathering Feedback for Teaching: Combining High-Quality Observations With Student Surveys and Achievement Gains," (Bill & Melinda Gates Foundation, January 2012) at [http://www.metproject.org/downloads/MET\\_Gathering\\_Feedback\\_Practitioner\\_Brief.pdf](http://www.metproject.org/downloads/MET_Gathering_Feedback_Practitioner_Brief.pdf).

## Additional Sources

Betebenner, Damian W., "Norm- and Criterion-Referenced Student Growth," *Educational Measurement: Issues and Practice*, 28(4) (2009): 42–51.

Betebenner, Damian W. and Linn, Robert L. "Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability," ETS Exploratory Seminar Research Monograph. (Educational Testing Service, 2010).

Briggs, Derek C. "The Goals and Uses of Value-Added Models." Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, D.C., November 2008.

Chetty, Raj, Friedman, John N., and Rockoff, Jonah. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." National Bureau of Economic Research working paper no. 17699. 2011.

Easton, John. "Goals and Aims of Value-Added Modeling: A Chicago Perspective." Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC., November 2008.

Glazerman, Steven, Loeb, Susanna, Goldhaber, Dan, Staiger, Douglas, Raudenbush, Stephen, and Whitehurst, Grover. "Evaluating Teachers: The Important Role of Value-Added." Washington, D.C.: Brown Center on Education Policy at the Brookings Institution, 2010, <http://www.brookings.edu/research/reports/2010/11/17-evaluating-teachers>.

Goldhaber, Dan, Gabele, Brian, and Walch, Joe. "Does the Model Matter? Exploring the Relationship Between Different Achievement-based Teacher Assessments." CEDR Working Paper 2012-6. University of Washington, Seattle, 2012.

Goldhaber, Dan, and Hansen, Michael. "Using Performance on the Job to Inform Teacher Tenure Decisions," *American Economic Review* 100(2) (2010): 250-255.

Goldschmidt, Pete, Roschewski, Pat, Choi, Kilchan, Auty, William, Hebbler, Steve, Blank, Rolf, and Williams, Andra. "Policymaker's Guide to Growth Models for School Accountability: How Do Accountability Models Differ?" Washington, DC: Council of Chief State School Officers, 2006.

Hanushek, Eric A. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *Journal of Human Resources*, 14(3) (1979): 351-388.

Harris, Douglas N., *Value-Added Measures in Education: What Every Educator Needs to Know*. Cambridge, Massachusetts: Harvard Education Press, 2011.

Kane, Thomas J., McCaffrey, Daniel, Miller, Trey, and Staiger, Douglas, "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignments" (Seattle, Wash.: Measures of Effective Teaching Project, Bill & Melinda Gates Foundation, 2013).

McCaffrey, Daniel F., Lockwood, J. R., Koretz, Daniel M., and Hamilton, Laura S. *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, California: RAND Corporation, 2003.

Rothstein, Jesse. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1) (2010): 175–214.

This publication features information from public and private organizations and links to additional information created by those organizations. Inclusion of this information does not constitute an endorsement by the U.S. Department of Education of any products or services offered or views expressed, nor does the Department of Education control its accuracy, relevance, timeliness or completeness.